893 Final Project

Jiaying Li & Tzu-Chi(Stephanie) Lin

Due May 8th, 2021

1 Introduction

The Human Connectome Project (HCP) is a five-year project launched in July 2009. The project aims to characterize human brain connectivity in about 1,200 healthy adults and to enable detailed comparisons between brain circuits, behavior, and genetics at the level of individual subjects. The human brain connectivity is explored using two complementary MR imaging modalities: diffusion imaging and resting-state fMRI.

The diffusion imaging is used to chart the trajectories of fiber bundles coursing throughout the brain's white matter. This is being done using HARDI (High Angular Resolution Diffusion Imaging) to acquire the data and probabilistic tractography to estimate fiber trajectories and generate maps of structural connectivity between gray matter regions. The resting-state fMRI (R-fMRI) is providing individual functional connectivity based on correlations in the fMRI BOLD signal among functionally interacting cortical and subcortical gray matter brain regions. Additional information about brain structure and function can also be obtained using Task fMRI, where subjects carry out a variety of behavioral tasks in the MR scanner. To understand the brain-behaviour relations, human traits data are collected using behavioral testing to assess individual differences in sensory, motor, and cognitive function.

In this report, we are interested in studying the relationship between human brain connectivity and human traits. A brief overview of the dataset and details about the data cleaning process are given. Then an exploratory data analysis has been applied. Based on our findings in the exploratory analysis, we selected features to predict human traits and summarized our results in the final section.

2 Overview

In this study, we consider 1065 young adults brain network data from the Human Connectome Project. Note that each subject has two 68×68 structural connectome (SC) and functional connectome (FC) matrices. As these matrices are symmetric and their main diagonals have the same value, we only vectorize the upper off-diagonal entries. Notice that the vectorized SC matrix has 76 all zeros columns. We remove these columns as they are not contributing any meaningful information.

We also included tensor PC scores that summarized structural and functional brain network data. There are 60 tensor PC scores from structural connectome features and 60 from functional connectome features. For the human traits data, we consider 175 human traits that are previously identified to be related to the brain network data.

In the 175 traits data, we examine the number of missing values in each trait. Figure 1 shows the top 10 traits that have the most missing values. The y-axis represents the percentage of NAs and the x-axis give the corresponding trait ID. Observed that the top 10 traits with the most number of NAs are traits in the substance use category. The top 5 traits are related to tobacco use and dependence and have more than 60% NAs. Hence, we remove 6 substance use traits that have more than 40% missing data.

Then, we merge the vectorized SC, vectorized FC, TNPCA data, and 169 traits based on subjects. We removed 7 subjects without the vectorized FC information. After removing these 7 subjects, Trait "EVA_Num" becomes a constant vector, which has also been removed. In the end, we are left with 168 human traits. Among all human traits, 138 are continuous data, 22 are ordinals, and 8 are binary data. Table 1 shows the number of traits in each category.

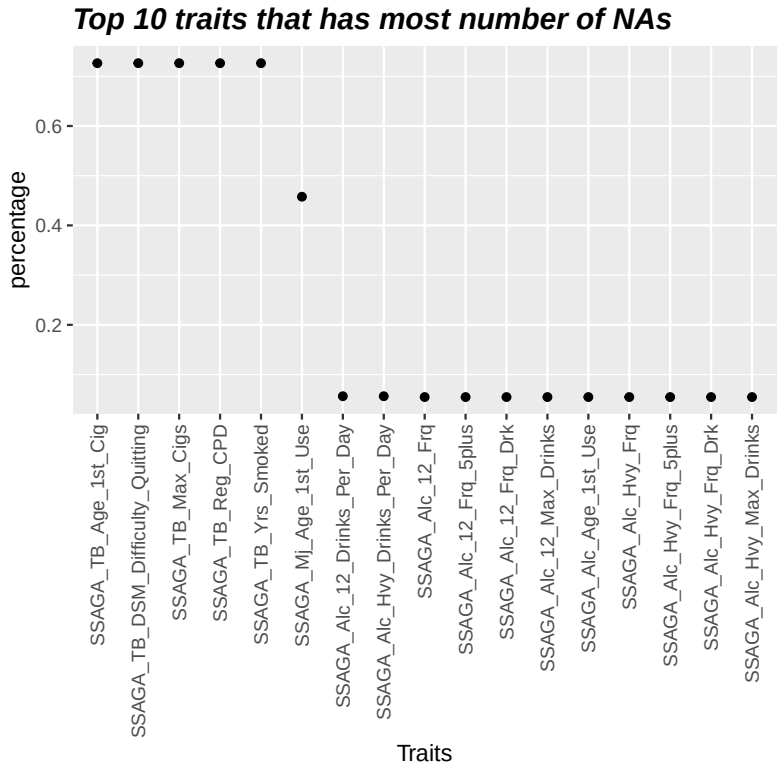
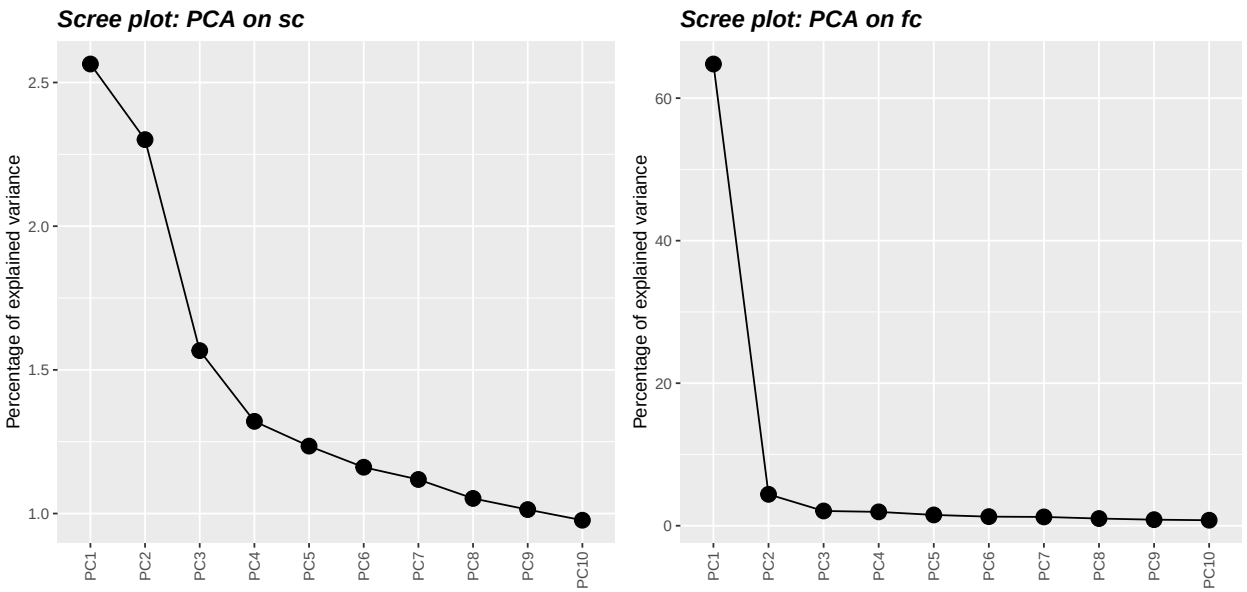


Figure 1: Top 10 traits that have the most NAs

Category	Alertness	Cognition	Emotion	Health & Family History	Motor
# of traits	1	45	24	3	7
Category	Personality	Psychiatric & Life function	Sensory	Substance Use	
# of traits	5	43	11	30	

Table 1: Traits Category

Observe that each subject contains 2202 vectorized SC features and 2278 vectorized FC features. To reduce the dimension for network data, principal component analysis (PCA) is applied to vectorized SC and FC data. Figure 2 shows the scree plots for SC and FC. The x-axis represents the first 10 principal components, while the y-axis represents the percentage of explained variance. In figure 2a, we observe that the percentage of explained variance for each SC principal component is below 3%. The first 161 principal components explain 60 % of the data. As of figure 2b, we observe that the first principal component for FC explains over 60% of the data.



(a) (b)

Figure 2: Scree plots for SC and FC

Below is a brief summary of the data management process

- For each subject, vectorize the structural connectomes (SC) and funcational connectomes (FC).
- Remove 76 all zeros columns in vectorized SC.
- Remove 6 substance use traits that contain more than 40% NAs.
- Merge vectorized SC, vectorized FC, TNPCA data, and 169 traits based on subjects.
- Remove 7 subjects without FC information.
- Remove trait that contains only one constant value: "EVA_Num".
- Apply PCA to vectorized SC and FC.

After the data management process, we have 1058 subjects and each subject contains SC, FC, TNPCA data, and 168 traits. In the merged data, the remaining 168 traits still have some missing values. If removing all subjects containing NAs, we are left with 890 subjects. To preserve as many observations as possible, we remove subjects containing NAs based on the traits examined.

3 Exploratory Data Analysis

In the following analysis, we focus our interests on human traits with continuous values and explore their relationship with brain network data. Figure 3 shows the number of continuous traits in each category. There are nine different categories, where cognition and the psychiatric & life function human traits are the two largest categories.

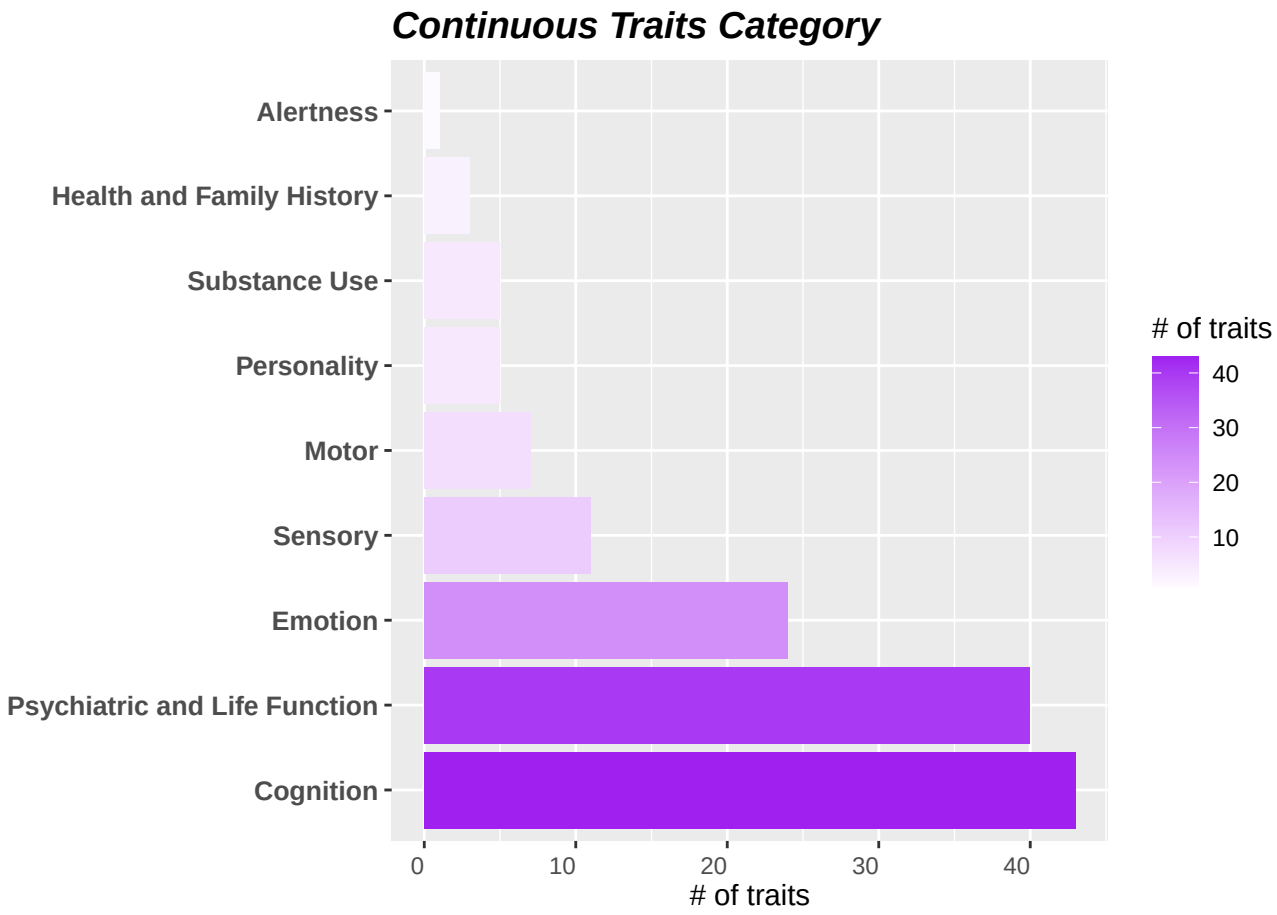


Figure 3: Number of continuous traits in each category

We first examine the linear relationship between the first PC score for brain network data and continuous traits shown in figure 4. All four figures have the x-axis represent the i-th continuous trait and the y-axis represents the correlation coefficient. The red dashed lines are for correlation at 0.3 and -0.3. On the upper left of figure 4 shows compare the first PC score for SC with continuous traits, while on the upper right of figure 4 shows the correlation between the first PC score for FC and continuous traits. The bottom two figures of 4 considered the first tensor PC scores for SC and FC.

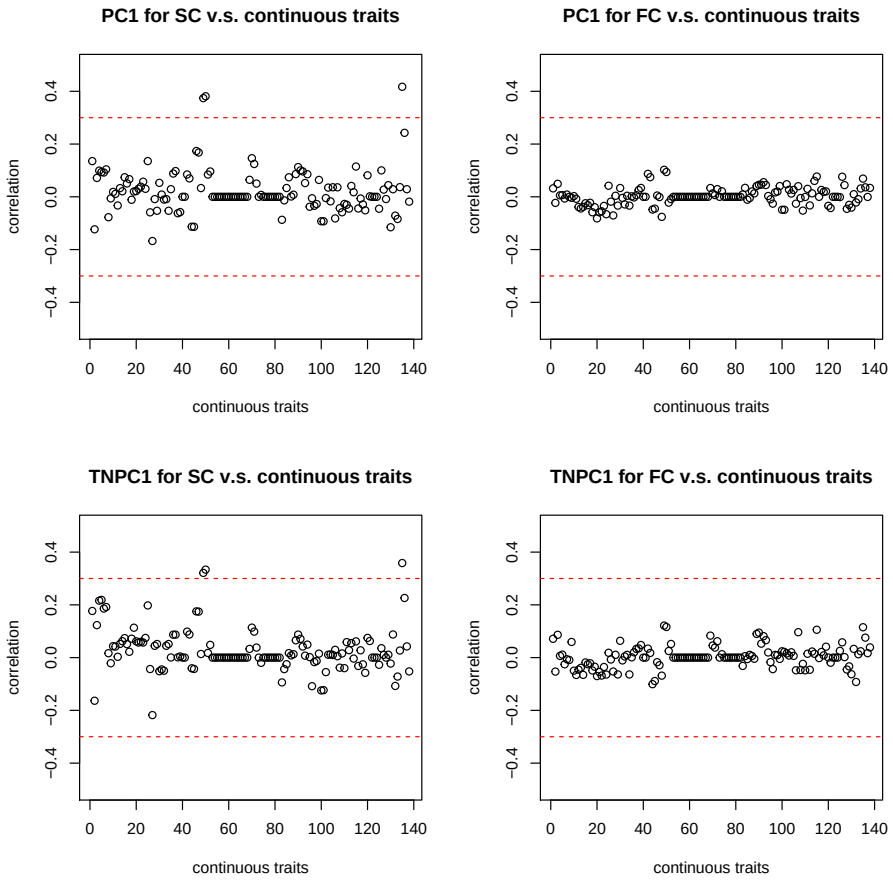
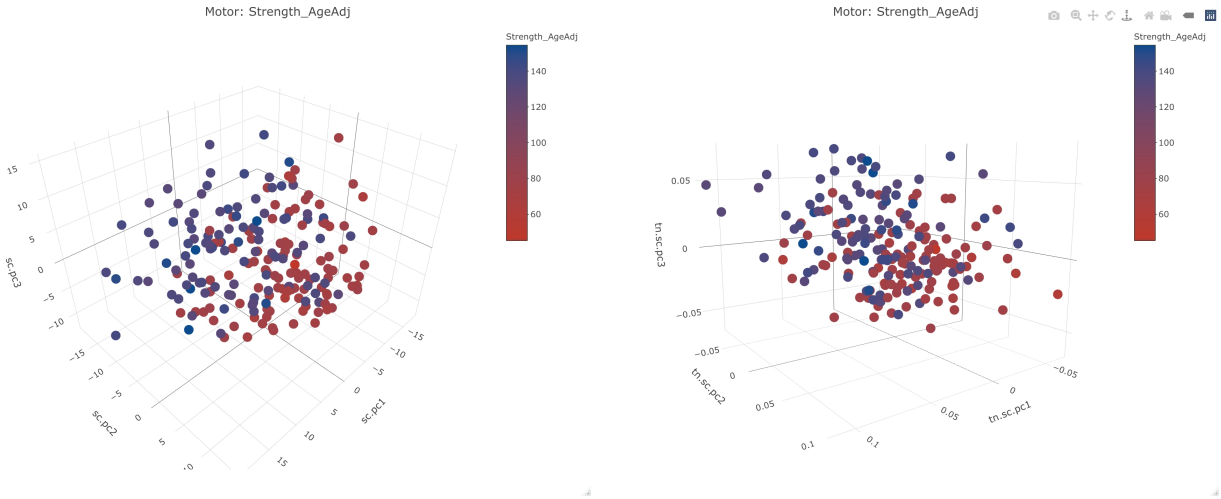


Figure 4: Correlation plots for the first PC score on brain network data and continuous traits

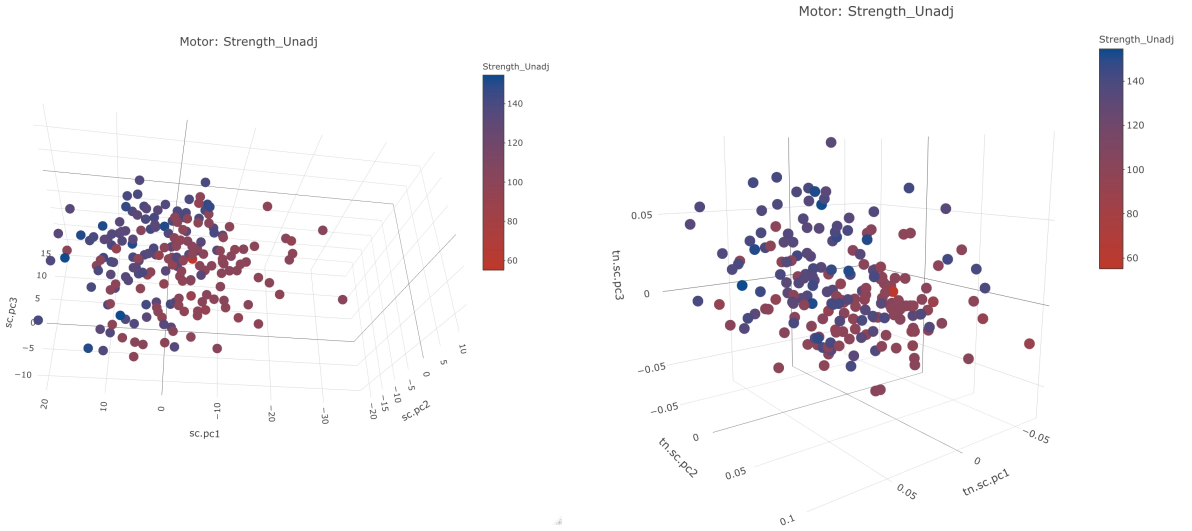
Observe that all four figures have the correlation coefficients fell between -0.3 and 0.3, which indicates that most continuous traits are weakly correlated to first PC scores for brain network data. The continuous traits "Strength_Unadj", "Strength_AgeAdj", and "Height" are moderately correlated to the first PC score for structural-related brain network data. Traits "Strength_Unadj", "Strength_AgeAdj" are in the motor domain, while "Height" is in the health and family history domain. The correlation coefficients suggest that traits of continuous type and the first PC score for brain network data may not be linearly dependent. Also, human traits in the motor domain seem to associate with structural-related brain network data.

Next, we examine the relationship between structurally related brain network data and two continuous traits "Strength_Unadj", "Strength_AgeAdj" in the motor domain. We first selected 100 subjects with low values in traits and 100 subjects with high values in traits. Then we display the first three PC scores for structural related brain network data along with two selected traits in figure 5a, 6a, 5b, and 6b. We can observe a separation between different groups of subjects from all figures.



(a) First 3 PC scores on SC with Strength_AgeAdj (b) First 3 tensor PC scores on SC with Strength_AgeAdj

Figure 6: First 3 PC scores on structural related network data v.s. motor trait: Strength_AgeAdj



(a) First 3 PC scores on SC with Strength_Unadj

(b) First 3 tensor PC scores on SC with Strength_Unadj

Figure 5: First 3 PC scores on structural related network data v.s. motor trait: Strength_Unadj

Also, we selected 100 subjects with low values in traits and 100 subjects with high values in traits and then apply two-sided hypothesis testing to their first PCA scores on SC and FC. The results are shown in figure 7. The x-axis represents the corresponding category for continuous traits, while the y-axis represents the p-values. The red dashed line shows the threshold for the p-value smaller than 0.05. From the hypothesis testing results, the structural network data could discriminate traits in the motor domain well. Also, we observe that the first PC score for FC seems to separate high and low values for sensory and substance use

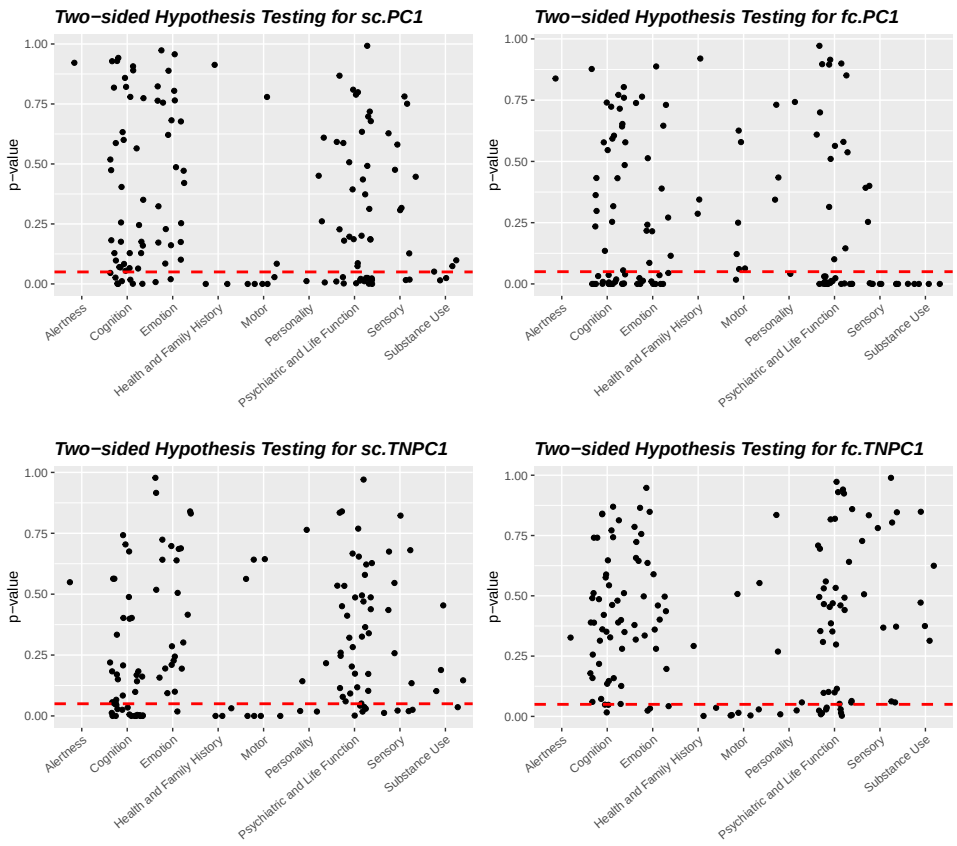


Figure 7: Two-sided hypothesis testing results to continuous traits

traits. It seems that structural and functional brain networks can be used to distinguish high and low values in some continuous traits.

4 Predictive Analysis

In this section, we will study the relationship between brain connectivity and traits. Let's see whether brain connectivity can predict traits.

We first try to predict "StrengthUnadj" since we know from exploratory data analysis that "Strength_Unadj" has a relatively high correlation with first PC score for structural-related brain network data. We expect that our methods will have a good performance on predicting "StrengthUnadj". We will predict other traits later.

We use SC, FC and TNPCA data to make prediction. To reduce dimension, we don't use all SC and FC data. First 161 columns of SC data can explain at least 60 % of variance in SC data while first 160 columns

can not. So we only keep first 161 columns in SC data. 1st column of FC data can explain more than 60 % of variance in FC data so we only keep 1st column of FC data. Now our predictor includes 161 columns of SC data, 1 column of FC data, 120 columns of TNPCA data.

4.1 Regression

Category of "StrengthUnadj" is continuous. We will try some regression methods. We use linear regression, single tree and randomforest to make prediction, here linear regression can be viewed as benchmark.

After removing subjectives that have NA in "StrengthUnadj", we have 1057 subjects. Then we split the data into training data and testing data. Approximately 2/3 of data is training dataset and remaining 1/3 of data is testing data.

Then we perform linear regression. Multiple R-squared is 0.6391, adjusted R-squared is 0.3979. Testing error is 11.51053. From R-squared values, we know our predictor can predict "StrengthUnadj" to some extent when using the simplest linear regression model. Let's see whether single tree and randomforest will have a better predictive power for "StrengthUnadj".

Testing error of single tree is 11.64323. It's a bit larger than the testing error of linear regression. So we can say single tree doesn't have a better predictive power.

We want to build the best possible random forest, so we need to find optimal parameters *mtry* and *ntree* in the model.

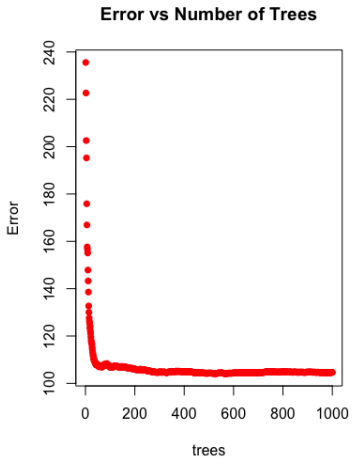


Figure 8: Error vs. *ntree*

We first set $mtry = 10$ and $ntree = 1000$. From Figure 8, we see error doesn't decrease when the number of

trees is larger than 300. So 300 trees is enough. We use $ntree = 300$ in our final model. Then we need to decide $mtry$.

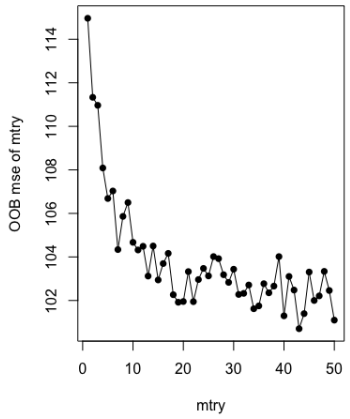


Figure 9: Error vs. $mtry$

From Figure 9, we choose $mtry$=43 since it minimizes error. In our final model, $ntree$=300 and $mtry$=43. Testing error is 10.34742. Randomforest slightly improves predictive power compared to linear regression and singe tree.

4.2 Classfication

We can also do classification for "StrengthUnadj" although it's a continuous variable. We reorder the data based on "StrengthUnadj" value than we create high/low (1/0) label. 50 % of data is labeled as high while the remaining 50 % is labeled as low. Then we scale the predictors before building our model. We will consider KNN and SVM in classification.

Square root of number of observations in training data is around 26.55, so we create two models. One with K value as 26 and the other model with K value as 27. Classification accuracy for these two models are 61 % and 60% respectively.

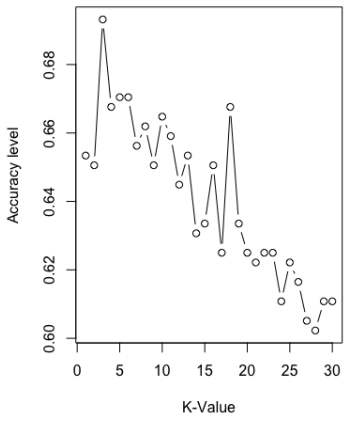


Figure 10: KNN

From Figure 10, we know when k=3, KNN model has the highest classificaiton accuracy 69 %.

Then we try SVM. When we use linear kernel, classification accuracy is 66%. When we use radial kernel, classification accuracy is 73%.

SVM with radial kernel has the highest classification accuracy 73% for "StrengthUnadj".

4.3 Prediction results

We have tried three regression methods (linear regression, single tree, randomforest) and two classification methods (KNN, SVM) for "StrengthUnadj". Now we make prediction on all 138 continuous traits.

We have testing error for linear regression model, single tree and randomforest. We also have classification accuracy for KNN, SVM with linear kernel and SVM with radial kernel.

Table 2 and Table 3 show part of our results.

Testing error of single tree and randomforest are almost the same as testing error of linear regression. Single tree and randomforest don't improve predictive power.

SVM with radial kernel performs better than SVM with linear kernel. For some continous traits, SVM with radial kernel has a higher classification accuracy than KNN, for other continous traits, KNN has a better performance.

We have 138 continous traits. For 135 of them, both KNN and SVM (with radial kernel) have a classification

Trait	Linear regression	Single tree	Randomforest
Edurance_UnAdj	15.5425501	14.0776366	13.1588653
Edurance_AgeAdj	18.2140049	17.4669208	15.3036013
GaitSpeed_Comp	0.2900603	0.2508767	0.2276338
Strength_Unadj	11.5105328	11.6432316	10.3489604
Strength_AgeAdj	20.3156004	21.0215660	18.6210587

Table 2: Testing errors using SC and FC data

Trait	KNN	SVM (linear)	SVM (radial)
Edurance_UnAdj	0.5880682	0.5454545	0.5965909
Edurance_AgeAdj	0.6335227	0.5511364	0.6051136
GaitSpeed_Comp	0.5977337	0.5920680	0.5779037
Strength_Unadj	0.6051136	0.6562500	0.7272727
Strength_AgeAdj	0.6789773	0.6590909	0.7159091

Table 3: Classification accuracy using SC and FC data

accuracy lower than 70 %. The remaining three continous traits are Strength_Unadj, Strength_AgeAdj and Height. For these three traits, SVM has classification accuracy higher than 70 % while KNN has classification accuracy lower than 70 %.

Note that we mentioned in previous section that "StrengthUnadj", "StrengthAgeAdj", and "Height" are moderately correlated to the first PC score for structural-related brain network data. So we would like to see whether SVM still have good classification power if we only use SC data. The results are shown in Table 4.

Classification accuracy doesn't change much when we only use SC data instead of using SC data and FC data. So the reason that SVM has good predictive power for these three traits might be the relatively high correlation between first PC score for structural-related brain network data and trait.

Trait	SVM (SC and FC data)	SVM (SC data)
Strength_Unadj	0.7272727	0.6988636
Strength_AgeAdj	0.7159091	0.7357955
Height	0.7670455	0.7585227

Table 4: Classification accuracy comparison

5 Summary

In this report, we focus on studying the relationship between brain network data and human traits of continuous type. In the exploratory data analysis, we observe that most human traits and brain network, either structural or functional-related, may not be linear dependent. In addition, for motor behaviour, we observed structural connectomes and grip force are positive correlated. We also find structural connectomes and height are positive correlated. Based on the findings in EDA, it seems that structural-related brain network data can explain variation in "StrengthUnadj", "StrengthAgeAdj", and "Height" using nonlinear model.

In prediction analysis, we find that testing error of single tree and random forest are almost same as testing error of linear regression. Single tree and random forest doesn't have better predictive power than linear regression. SVM with radial kernel has a higher classification accuracy than SVM with linear kernel so the relationship between brain network data and traits might be nonlinear. SVM(with radial kernel) has a high classification accuracy (higher than 70 %) on traits "StrengthUnadj", "StrengthAgeAdj" and "Height". Besides, classification accuracy on above three traits almost doesn't change when we use structural-related data instead of using structural-related and functional-related data. So these three traits confirms our findings in EDA that they are associated to the structural-related brain network data.

6 Authors' contribution

The authors confirm contribution to the project as follows: **Jiaying Li:** Vectorized network data and merge brain network data and human traits based on subjects. Conduct predictive analysis and interpretation of results. **Stephanie Lin**: Deal with missing data and perform PCA to network data. Perform exploratory data analysis and interpretation of results. All authors discussed and determined the project goal.